Statistics Iceland

Informed society

SDC Guidelines for Research Outputs dissemination and/or publication.

1. Introduction

Purpose and scope

The purpose of this manual is to assist researchers collaborating with Statistics Iceland's Research Services in ensuring that any data exported from the Data Portal, or published as part of research dissemination, maintains confidentiality and adheres to disclosure control standards.

This manual focuses on presenting and explaining key elements relevant to statistical disclosure control strategies. Although general concepts and measures are presented, an emphasis is placed on risk disclosure of data belonging to the research databases available through Statistics Iceland’s research services database.

Understanding disclosure risk

Disclosure occurs when “a person or organisation recognises or learns something that they did not know already about another person or organisation, via released data”. Statistical disclosure may occur when the publication of statistical information enables external users to identify and link confidential or sensitive data, either on its own or in combination with other publicly available sources.

Types of disclosure

Identity disclosure: It occurs when the identity of an individual or organization can be linked to a released data record that contains confidential or sensitive information.

Attribute disclosure: It involves (1) the uncovering of new confidential or sensitive information about an individual or organisation through the use of published data, (2) the uncovering of confidential or sensitive information about a group of identifiable individuals or organisations (a.k.a. group attribute disclosure).

Statistical Disclosure Control

Statistical disclosure control (SDC) techniques refer to a set of methods employed to reduce the risk of disclosing information on individuals or organisations. Statistics Iceland research services users are granted access to microdata for research purposes.

In regards of Statistics Iceland research services database, all data tables created for each research project are pseudo anonymised, by creating a random id for each statistical unit included. This id is unique for each project, ensuring that sensitive data cannot be liked to records in other projects the researcher might have access to.

2. Data types

Individual-level microdata

It usually focuses on individual persons but can also include households or be structured hierarchically with individuals within households. Individual-level data typically includes variables related to demographics (age, gender, marital status, education level, school attendance), income, expenditures, labour status, health, religion, or political affiliation. Variables can be categorical (e.g., gender, marital status) or continuous (e.g., age, income).

Firm-level microdata

It focuses on businesses, economic entities, enterprises, or companies. Firm-level data typically includes variables related to economic activity, size, location, turnover, sales, number of employees, investments, or profit. Variables in enterprise data often take the form of quantitative variables with asymmetric distributions.

Sensitive variables

The categorisation of a particular variable as sensitive can be context-dependent and level of aggregation. A non-exhaustive list of variables that are considered sensitive and might lead to identity of attribute disclosure includes:

- Identifiers or quasi-identifiers when linked to additional/external data
- Demographic characteristics (Individuals)
    - Age, sex, gender identity, country or birth
    - Citizenship
    - Marital Status
    - Household characteristics
- Socio-economic status (Individuals)
    - Income and earnings
    - Tax registry
    - Occupation or industry of employment
    - Education
    - Employment status
- Geographic information (Individuals)
    - Address
    - Municipality
- Health related data (Individuals)
    - Health status
    - Medical conditions and treatment received
    - Disability status
    - Parental leave
- Firm data (Organisations)

- Net turnover, value added, income, and profit
- Investment figures
- Number of employees/members/students and their characteristics

3. SDC for specific research outputs

The main warning-signals for the need of SDC-checks are (1) the presence of skewed distributions of the variables included/presented, and (2) the presence of sensitive variables.

List of output types to be protected

- Tabular outputs
  - Magnitude tables
  - Frequency tables
- Descriptive statistics
  - Means, indices, rations, indicators
  - Maxima, minima, and percentiles
  - Graphs and plots
- Data analysis results
  - Estimation residuals
- Code scripts

Other outputs that are not likely to be disclosive are:

- Complex descriptive statistics (e.g. mode of distributions, higher moments of distributions, concentration ratios)
- Analysis results
  - Coefficients of statistical models
  - Summary and test statistics from estimates
  - Correlation coefficients
  - Factor analysis
  - Correspondence analysis

Test and thresholds

The general types of tests described in this section refer to the number of units, degrees of freedom of modelled output, group attributes protection, and dominance rules. In this manual, thresholds are presented as ranges according to Eurostat guidelines. If needed, a random/confidential number(s) in the defined range will be assigned to the project based on the project's needs and intended outcomes' formats.

T1. Test and thresholds for tabular outputs (Frequency tables)

-T1.1. The number of statistical units (unweighted) for each cell is at least n, with n between 5 and 10.

-T1.2. No cell contains more than r% of the total number of units in its row or column, with the percentage r% between 85% and 95%.

- T1.3. No cell in the table is generated by units which belong to a unique source (sub-group/organisation) in proportion of more than a recommended r%, with the percentage r% between 85% and 95%.

-T1.E.1 Exceptions: Structural zeros, i.e. when logically certain counts cannot be other than zero. For example, impossible to have the highest level of education for the age group 0-4 years.

-T1.E.2 Exceptions: Depending on the projects' scope, the lower threshold (referring to T1.1) could be lower up to 3 for firm data, if microdata is not linked to external data imported by the researcher.

T2. Test and thresholds for tabular outputs (Magnitude tables)

T2.1. The number of contributors to a cell is at least N, with N between (5 and 10)

T2.2. The (n,k) dominance rule: A cell is flagged as sensitive if the sum of the values of the largest n respondents in that cell constitutes k% or more to the total value of the cell. Values of k for, n=1 and n=2, would be between 85%-97% and 88%-97%, respectively.

T2.3. The p% rule: A cell is flagged as sensitive if the value of the largest cell contributor can be estimated within p% of its value by knowing the values of the other, smaller, contributors. The value of p would be between 5% and 15%. This rule is methodologically recommended over the (n,k) rule.

T2.E.1 Exceptions: These thresholds values would be adjusted for data presented as time series. In this regard, a time point (or period) is sensitive if n or fewer contributions account for k% or more of the total value. Specifically, regarding dominance rules, the value of k for n=1 could be lowered up to 50%, depending on the period length, aggregation level, and the variable risk level.

-T2.E.2 Exceptions: Depending on the projects' scope, the lower threshold (referring to T2.1) could be lower up to 3 for firm data, if microdata is not linked to external data imported by the researcher.

T3. Test and thresholds for descriptive statistics

T3.1. Means, indices, ratios, and indicators:

a) Each single value should derive from the synthesis of at least n units, with n between 5 and 10.
b) For each value to be released, the largest contributor included in the synthesis cannot exceed 50% of the total

T3.2. Minima and maxima: Should only be published if they represent at least n units, as long as it does not pose group disclosure danger or does not point to an extreme unit in the sample studied. The value of n should be between 5 and 10.

T3.3. Percentiles: Should not be published if the rank ordering or the statistical units is known. Extreme percentiles should be capped, rounded, or reported in intervals if they pose group disclosure danger or point to an extreme unit. No percentiles should be reported if they represent less than n units.

T3.4. Plots: General rules for plots and graphs publication include:

a) Data points cannot be identified with units. Does not apply to graphs of transformed or fitted data.
b) They should not include significant outliers.
c) The scales should be adequately selected to prevent the identification of (nearly) exact values.
d) Graphs should be exported/published as .jpeg, .jpg, bmp, or .wmf files. With no data attached.

-T3.E.1 Exceptions: Depending on the projects' scope, the lower threshold (referring to T3.1a) could be lower up to 3 for firm data, if microdata is not linked to external data imported by the researcher.

T4. Rules for analysis results.

T4.1 Modelling residuals: Should be published only if they are needed. If that is the case, it should be done descriptively, following the rules for plots described above. More detailed information may be shared only for standardised residuals of aggregated figures and when the cells represent a minimum of n units, with n between 5 and 10.

T4.2 Modelling coefficients: They are considered low risk as long as the training data is large enough and the variables included are both categorical and numerical. For publication it is recommended to withhold at least one coefficient to lower the risk of reverse-engineering.

T5. Rules for code publication.

T5.1. Code files should be only published if they do not contain embedded disclosive data or insights about the dataset. These include hard-coded data, researchers' comments about the sample, and code that discloses information about a small sample.

4. Support and resources

Handbook on Statistical Disclosure Control © 2024 by Anco Hundepool, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Rainer Lenz, Jane Naylor, Eric Schulte Nordholt, Giovanni Seri, Peter-Paul De Wolf, Reinhard Tent, Andrzej Młodak, Johannes Gussenbauer, Kamil Wilak. URL: Handbook on Statistical Disclosure Control

Bond S., Brandt M., and de Wolf P-P.: Guidelines for the checking of output based on microdata research. Technical report, ONS, DeStatis, CBS, 2013. Project No: 262608. Data without Boundaries. WORK PACKAGE 11 (Improved Methodologies for Managing Risks of Access to Detailed OS Data). D11.8 - Final reports of synthetic data CTA, ECTA, cell suppression & Guidelines for output checking. URL: Output-checking-guidelines.pdf

5. Appendices

Glossary of terms

i. Microdata: Sets of records containing information on individual persons, households, or business entities. It is unit-level data, meaning each record represents a single respondent, such as an individual person or an economic entity.

ii. Sensitive variables: Variables within a dataset that, if known, could potentially lead to the identification of a statistical unit (e.g. individuals or organisations) and/or the disclosure of information considered private or confidential.

iii. Statistical disclosure control methods: Methods to reduce the risk of disclosing information on the statistical units.